Pros and cons of data sharing

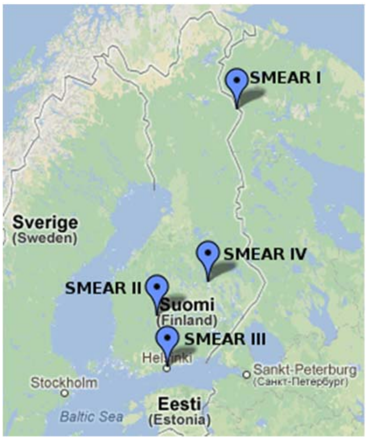
Open Forest Science 7.5.2018

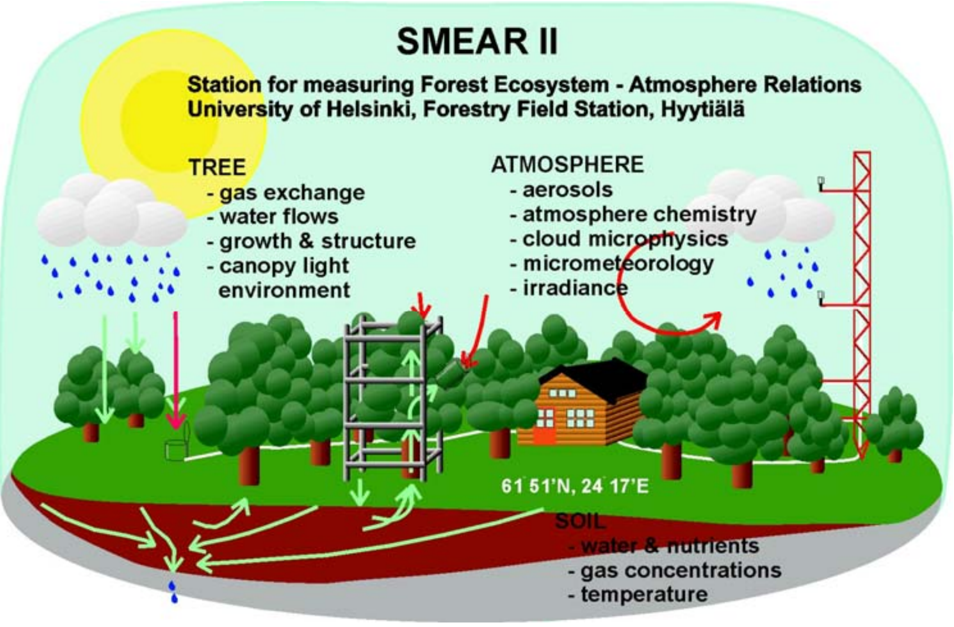
Pasi Kolari
University of Helsinki
Institute for Atmospheric and Earth System Research

Thanks: Timo Vesala, Toprak Aslan, Ari Asmi

INAR/SMEAR data

- Observations
  - Continuous measurements at SMEAR stations, large number of environmental variables, mostly time series
  - Short-term campaigns at SMEAR stations or elsewhere, also some geospatial data
- Field & lab experiments
- Modelling results
- No sensitive data

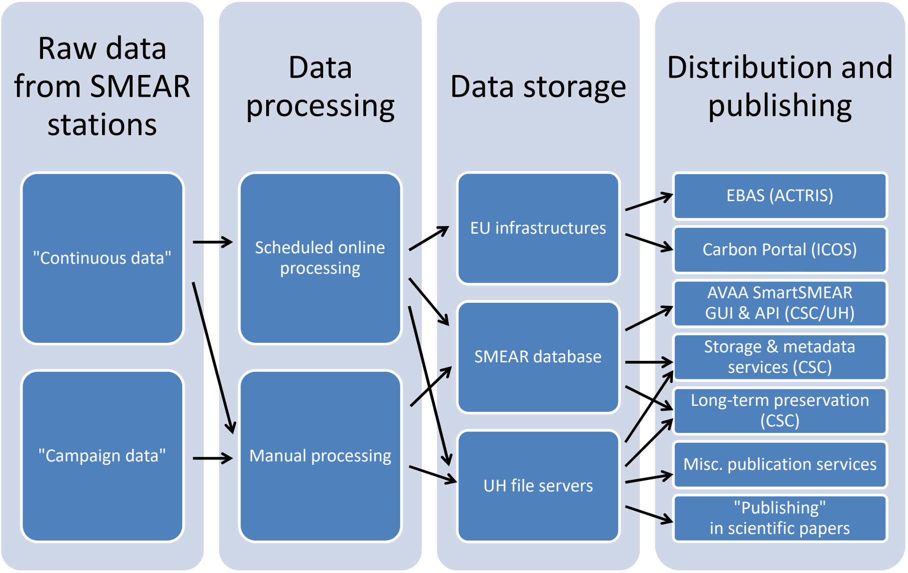




SMEAR II

Station for measuring Forest Ecosystem - Atmosphere Relations
University of Helsinki, Forestry Field Station, Hyytiälä

SMEAR data flow



Raw data from SMEAR stations

- "Continuous data"
- "Campaign data"

Data processing

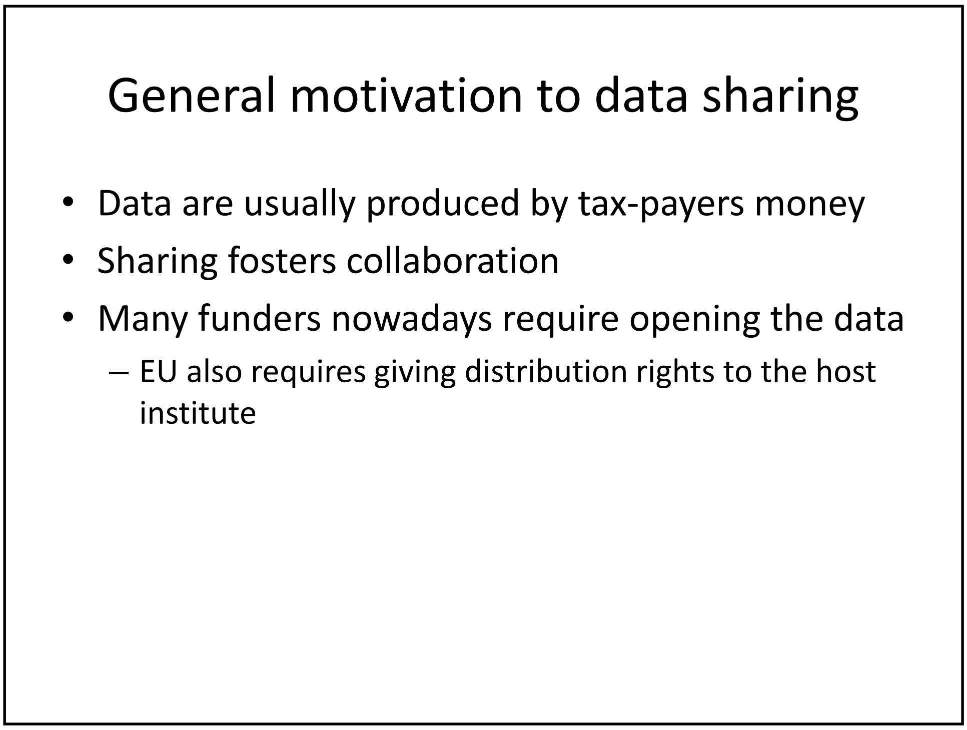
- Scheduled online processing
- Manual processing

Data storage

- EU infrastructures
- SMEAR database
- UH file servers

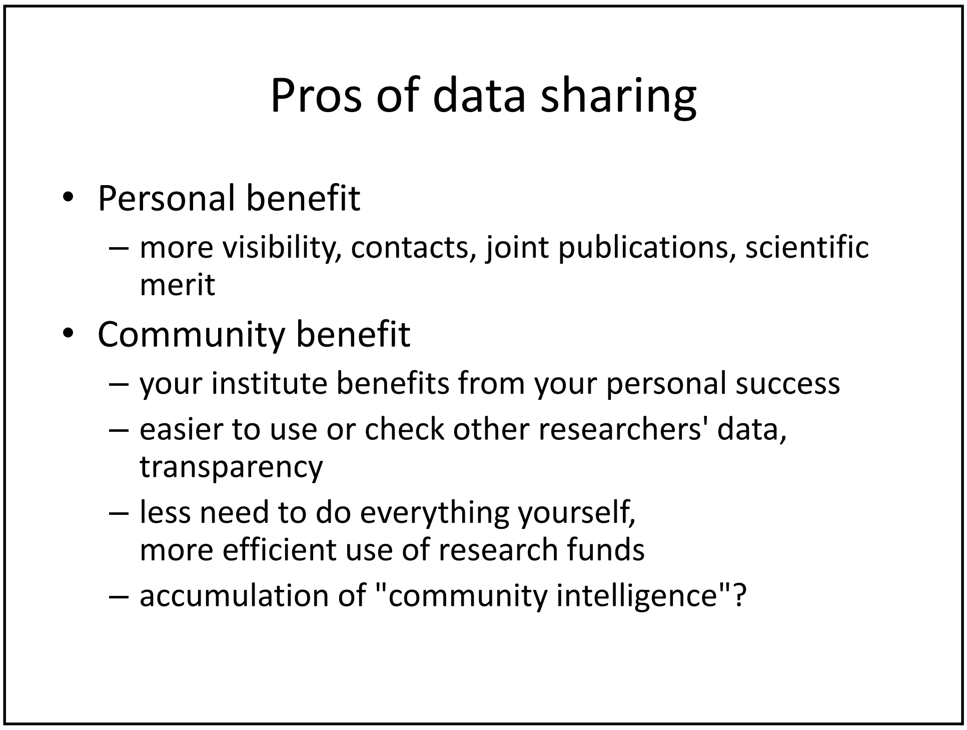
Distribution and publishing

- EBAS (ACTRIS)
- Carbon Portal (ICOS)
- AVAA SmartSMEAR GUI & API (CSC/UH)
- Storage & metadata services (CSC)
- Long-term preservation (CSC)
- Misc. publication services
- "Publishing" in scientific papers



General motivation to data sharing

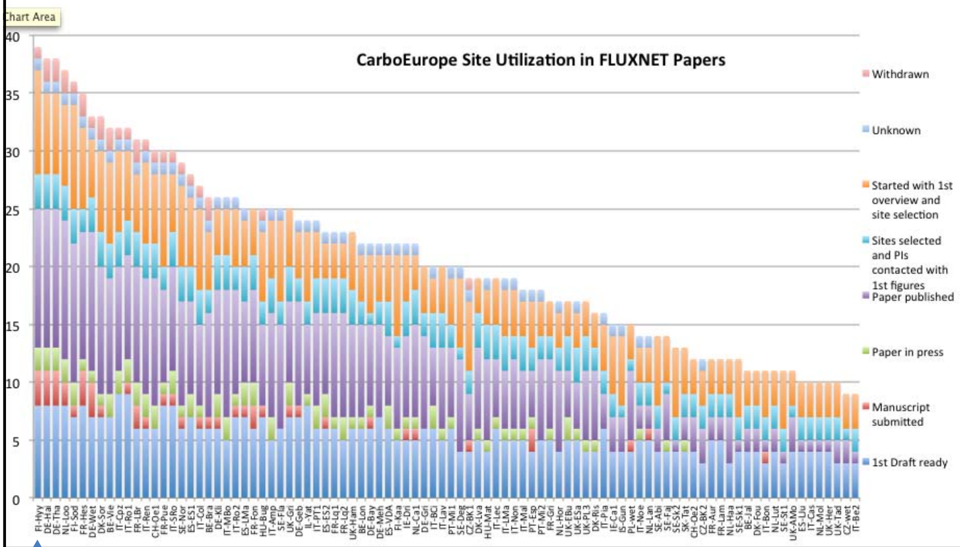
- Data are usually produced by tax-payers money
- Sharing fosters collaboration
- Many funders nowadays require opening the data
  - EU also requires giving distribution rights to the host institute



Pros of data sharing

- Personal benefit
  - more visibility, contacts, joint publications, scientific merit
- Community benefit
  - your institute benefits from your personal success
  - easier to use or check other researchers' data, transparency
  - less need to do everything yourself, more efficient use of research funds
  - accumulation of "community intelligence"?

How open data policy can be reflected in your publication list



↑

Hyytiälä (= station PI Timo Vesala)

Transparency of research

Letter

Nutrient availability as the key regulator of global forest carbon balance

M. Fernández-Martínez, S. Vicca, I. A. Janssens, J. Sardans, S. Luyssaert, M. Campioli, F. S. Chapin III, P. Ciais, Y. Malhi, M. Obersteiner, D. Papale, S. L. Piao, M. Reichstein, F. Rodà & J. Peñuelas

Nature Climate Change **4**, 471–476 (2014) doi:10.1038/nclimate2177	Received: 06 September 2013 Accepted: 28 February 2014 Published: 06 April 2014

Correspondence

Data quality and the role of nutrients in forest carbon-use efficiency

Werner L. Kutsch & Pasi Kolari

Figure 2 : The coupling between ecosystem respiration (Re) and gross primary production (GPP) weak in nutrient-rich forests and very strong in nutrient-poor forests.

From: Nutrient availability as the key regulator of global forest carbon balance

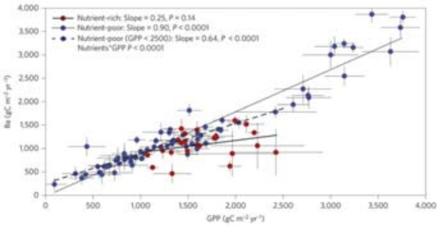
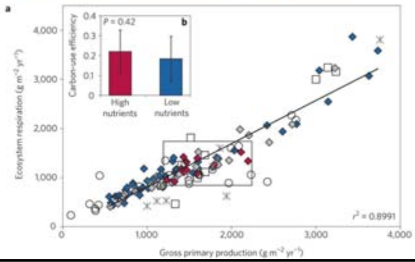


Figure 2 : Ecosystem respiration (Re) plotted against GPP for the remaining 82 sites.

From: Data quality and the role of nutrients in forest carbon-use efficiency



Are there drawbacks in data sharing?

- The idea of data sharing is not universal
  - Russia and China: just difficult to get data (competition, no sharing culture), although exceptions exist
- Sharing the data means additional work,
  where's the reward? Someone else is getting it?
  - Someone could "steal" your data or just use it before you can do it
- Data providers and end users tend to loose personal contact when everything is open and free
  - More chances for misinterpreting data

Additional workload from opening data

- Especially historical data are problematic
- The original documentation might be lacking or difficult to understand:

```
01:00 29/06/97-00:59 30/06/97
LIGHT POLE
measurements started at 23.35 PK
…
01:00 13/06/98-01:00 14/06/98
LIGHT POLE
Installed and started in a small configuration. E.S.
…
```

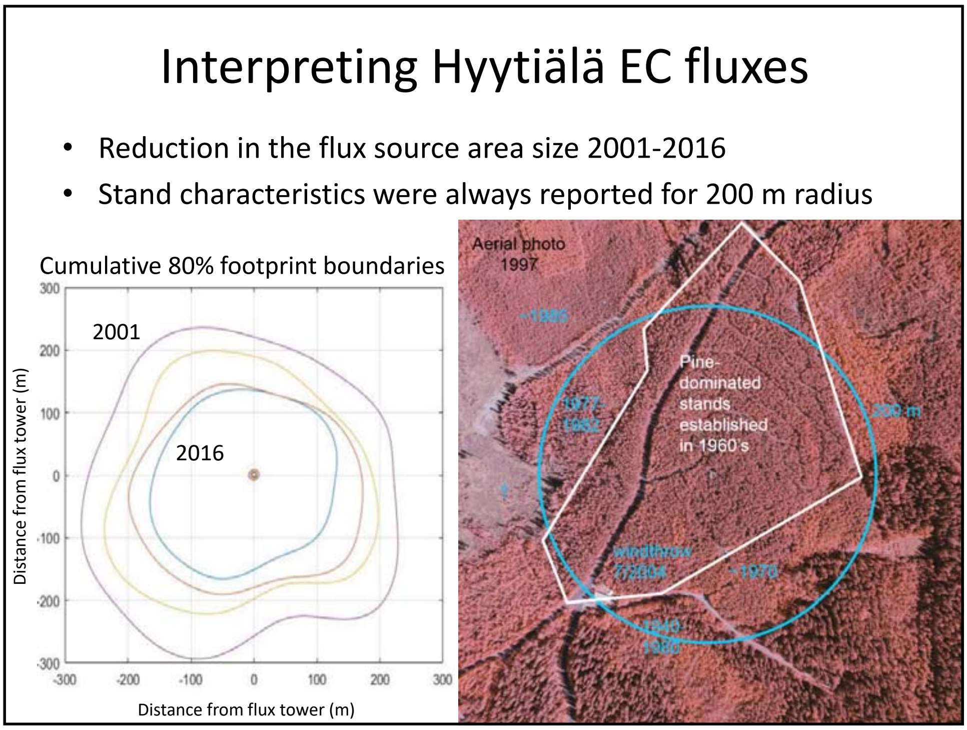
- Preservation of raw data from SMEAR stations for 50-100 years was considered too laborous compared to the foreseeable scientific value
  → only end-user data will be preserved "forever"

Reducing the documentation workload

- Be proactive!
- Start thinking about other users of data from the beginning of your project
- Find out what kind of metadata is needed when publishing the data
  - Publishing services usually provide GUI for metadata, no need to know much about metadata standards or xml
- Maintain systematic bookkeeping of measurements and data analyses, preferably in electronic form
  - This also benefits you personally
  - Not a huge task if started in time
- Prefer open file formats and keep your data tidy http://doi.org/10.5281/zenodo.400982

Data users have no contact with the original data provider

- Remote users of (field) data often make false interpretations
  - They might have little idea about the local environment
  - They might lack understanding on how instrument type or configuration affects the data
  - They might mix measured and gapfilled data
- Partial solution: Write better documentation!
  - One cannot document everything
- Never underestimate the ability of data users to misinterpret your data!



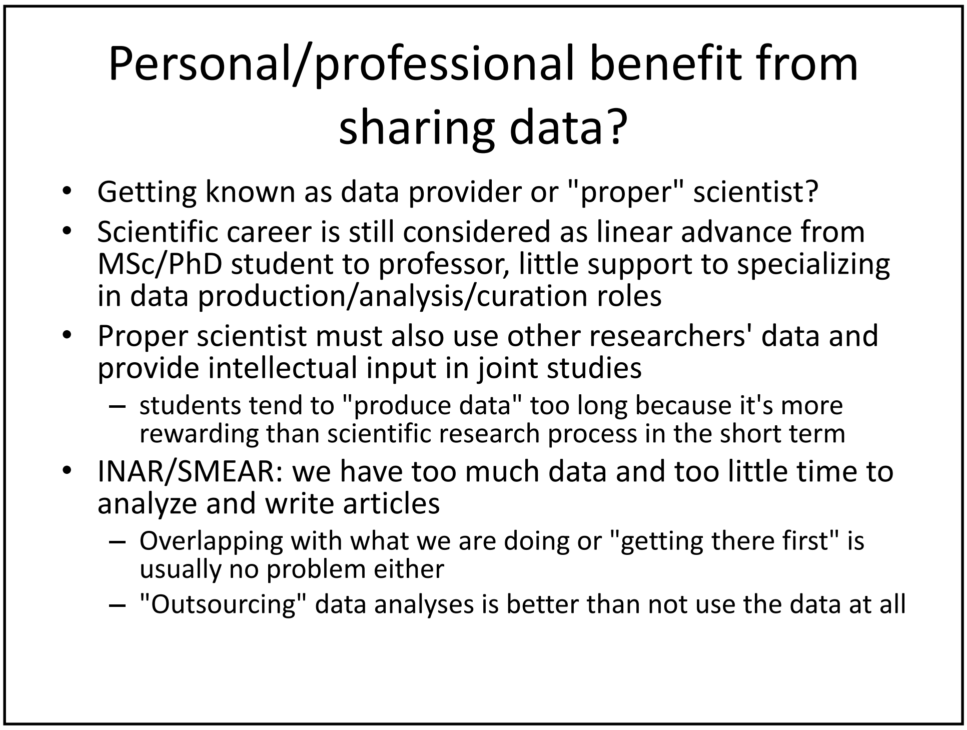
Interpreting Hyytiälä EC fluxes

- Reduction in the flux source area size 2001-2016
- Stand characteristics were always reported for 200 m radius

Cumulative 80% footprint boundaries

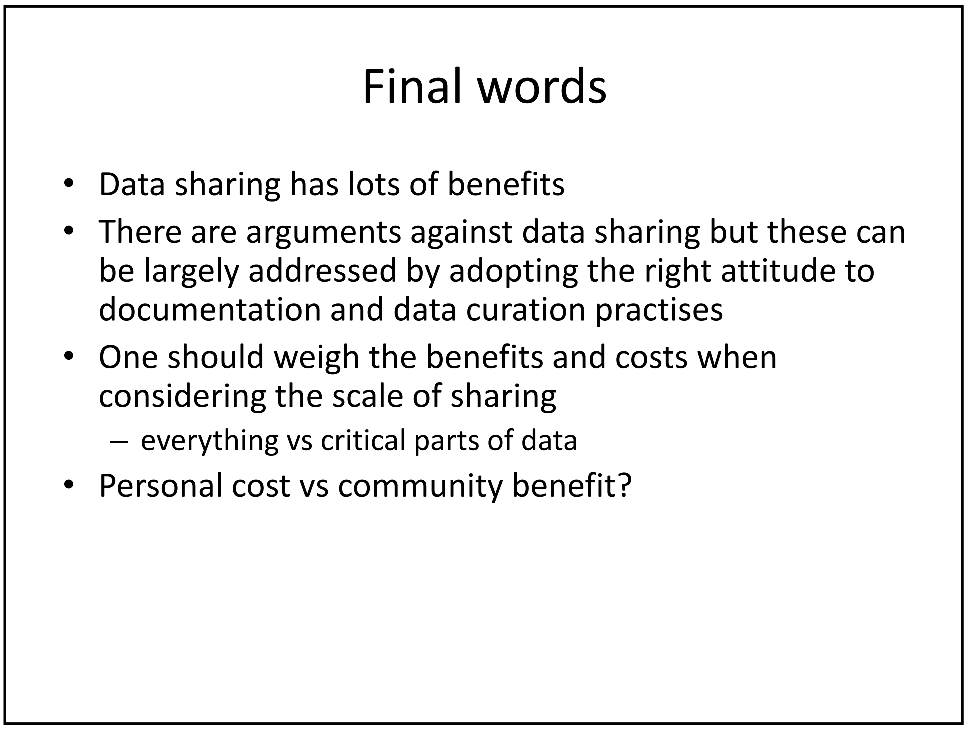
Reward from sharing your data?

- Acknowledgment practise from data production and use does not exist
  - Co-authorship or citing your scientific papers still the norm
  - Some publishers and institutes keep track of dataset references but there's no widely used system
- Publish your data via trusted public services that provide PID and harvestable metadata interface to ensure that your data has any chance to get registered in the future
- Publishing the data also ensures that you can claim your intellectual rights in the (rare) case that someone is using your data without acknowledgement
  - INAR/SMEAR: misuse of data is no problem (publications without co-authorship offers, citations, acknowledgements)
  - Getting caught from misusing someone's data is bad for your reputation



Personal/professional benefit from sharing data?

- Getting known as data provider or "proper" scientist?
- Scientific career is still considered as linear advance from MSc/PhD student to professor, little support to specializing in data production/analysis/curation roles
- Proper scientist must also use other researchers' data and provide intellectual input in joint studies
  - students tend to "produce data" too long because it's more rewarding than scientific research process in the short term
- INAR/SMEAR: we have too much data and too little time to analyze and write articles
  - Overlapping with what we are doing or "getting there first" is usually no problem either
  - "Outsourcing" data analyses is better than not use the data at all



Final words

- Data sharing has lots of benefits
- There are arguments against data sharing but these can be largely addressed by adopting the right attitude to documentation and data curation practises
- One should weigh the benefits and costs when considering the scale of sharing
  - everything vs critical parts of data
- Personal cost vs community benefit?